A neurocognitive framework for comparing linguistic and musical interaction

A neurocognitive framework for comparing linguistic and musical interaction

Lauren V. Hadley and Martin J. Pickering

University of Edinburgh

Running Head: Neurocognitive framework for interaction

Author Note

Lauren V. Hadley (corresponding author): Department of Psychology, University of Edinburgh, Edinburgh, EH8 9JZ. E-mail: lauren.hadley@cantab.net. Phone: +44 (0) 131 650 9867.

Martin J. Pickering: Department of Psychology, University of Edinburgh, Edinburgh, EH8 9JZ. E-mail martin.pickering@ed.ac.uk. Phone: +44 (0) 131 650 3447.

Abstract

Most cognitive research concerned with the relationship between language and music asks whether isolated individuals represent and process them in similar ways. In this paper, we focus instead on the relationship between interactive language and interactive music, and suggest that speakers engaged in dialogue and musicians engaged in joint performance face similar difficulties – how to relate their contributions to their partners in terms of both timing and content. We propose a model that spans interactive language and music in which each interactor constructs a single joint prediction of their own and their partner's behaviour, and then compares that prediction against the actual behaviour when it occurs. We discuss how predictions differ depending on turn organisation, as well as message spontaneity. We relate this proposal to behavioural and neuroscientific data from interaction research in the domains of both music and language.

Keywords

Simulation; Predictive processing; Music and language

Introduction

When considering natural interaction, the first activity to come to mind is conversation. However, most of us are also used to singing happy birthday with friends, or massacring pop duets in karaoke. Indeed both language and music are characteristic human abilities, sharing many apparent similarities in terms of their origins, functions, and organization. Just as rhythmic information is used to maintain an isochronous beat in music, it is used to establish a regular syllable rate in language. Just as pitch conveys the progression of a musical melody, it underpins the prosody of linguistic discourse. Given such similarities, we might expect them to be processed similarly, and indeed many researchers assume that this is the case (Fitch, 2006; Patel, 2003). But much prior work has focused on processing within an individual, looking at the passive listener or the isolated speaker or performer, as opposed to the interactive pair or group. We draw together research across language and music, taking advantage of the diversity of work focused on interpersonal coordination to identify cognitive processes involved in prediction during communicative interaction.

For decades in the psychological study of production and comprehension, it has been assumed that an individual hearing a recording in the lab is a proxy for a conversational listener, and an individual speaking disconnected sentences into a microphone is a proxy for their talkative partner (see Clark, 1996). The use of such 'isolation paradigms' (Becchio, Sartori, & Castiello, 2010) is based on the assumption that the way an individual represents their own and their interaction partner's part captures the entirety of social cognition (Fuchs & De Jaegher, 2009). But while theoretically two such individuals could approximate a conversing pair, such an approach removes the social aspect of interaction (i.e. the possibility of adaptation based on feedback), and cannot capture processes that emerge in the pair as a coupled unit. Recent developments in technology, coupled with a shift in perspective, has led to research exploring how individuals coordinate during the reciprocal process of sending and

receiving information (Schilbach et al., 2013), and particularly the neural correlates of such social coordination (Bögels, Barr, Garrod, & Kessler, 2015; Menenti, Pickering, & Garrod, 2012; Noordzij et al., 2009). Here we draw from such work to address interaction processes across two forms of communicative interaction: language and music.

We focus on interaction as a shared cooperative activity that involves (1) mutual responsiveness – i.e. each individual adapting to the other with the aim of coordinating, (2) commitment to the joint activity – i.e. each individual aiming to generate their own actions to mesh with those of the others, and (3) commitment to mutual support – i.e. willingness to support each other where necessary for interactive success (Bratman, 1992). The linguistic and musical interactions that involve shared cooperative activity can be located along two key dimensions. The first is turn organisation (consecutive or concurrent). In speech, consecutive turns are common, with average inter-turn intervals being around 250ms (Stivers et al., 2009). However, while the primary speaking role alternates in consecutive interaction, turns can overlap, and listeners frequently produce "backchannel" or supporting contributions during these turns (Clark, 1996). Concurrent turn production, on the other hand, occurs when individuals produce outputs of similar significance at the same time. In language, concurrent turn production primarily involves speaking the same words, but in music it can involve producing different (but related) outputs. People are able to synchronise both their speech and their music performance remarkably well during concurrent turn production (around 30ms, Cummins, 2003; Keller, Knoblich, & Repp, 2007).

The second dimension of linguistic and musical interaction is message spontaneity. Both language and music can involve messages being improvised (spontaneously generated) or scripted (pre-determined), but while most speech is spontaneous (with scripted utterances being limited to performers), in music the dominant format is genre-dependent (e.g. classical music is often scripted while jazz music is often improvised). Notably, scripted interactions

clearly define the content to be produced by each interactor in terms of specific words or notes, whereas improvised interactions afford interactors autonomy over content.

Forms of Interaction

Consecutive Language: Consecutive language can be improvised (i.e. in conversation) or scripted (e.g. between actors). The conversation in Figure 1 is a representative example of spontaneous consecutive language. Characteristic of natural dialogue, in which interpretation of each interlocutor's contribution depends on associated contributions by the partner, the conversation would not be possible to understand from either interlocutor's contributions alone. Content is instead constructed jointly, for example when J begins complimenting how C cleans their shoes, C then explains the cleaning technique to J. While the repetitive nature of their speech would be unnecessary in a monologue, conversation requires interactors to flexibly provide, and respond to, feedback. Hence in addition to the difficulties inherent in planning and producing speech, individuals must monitor other interlocutors' interruptions for indications of failures to understand, then use these indications to update their speech plans; after J signals confusion at (3), C makes a clarification at (4). J then provides a backchannel signal that they have understood at (5). Each individual must be responsive to the cues of their partner to ensure mutual understanding (see Clark, 1996).

1 - J: Those shoes look nice when you keep putting stuff on them

2 - C: Yeah I have to get another can cuz it ran out. I mean it's almost out

3 - J: Oh ah

4 - C: Yeah well it cleans them and keeps ⌜ them clean

5 - J: ⌞ Yeah right

Figure 1. Simplified conversation example from Sacks, Schegloff, & Jefferson (1974).

Consecutive Music: Consecutive music can either be scripted (i.e. precomposed) or improvised, and in either case the musical content can only be understood from the combined contributions of all musicians. Musical turn-taking is common in forms such as jazz, and involves many of the same issues as linguistic turn-taking. In such music, each individual's contribution is influenced by that of the other(s), both in terms of structure and melodic features. Notably, in consecutive music production a non-soloist musician may produce supporting output during the partner's turn, providing accompaniment somewhat comparably to a conversational listener's backchannelling.

Concurrent Language: While it is possible for individuals to produce different words simultaneously, language becomes very hard to understand when overlap is extensive. If two individuals (typically addressees) start to speak concurrently (e.g., both responding to a question or statement), one tends to yield the floor rapidly to the other (see Schegloff, 2000). Hence extended concurrent speech is unlikely to be improvised, and is typically limited to scripted contributions (e.g. chanting), in which case each individual's contribution is likely to be very similar. Since the speech content itself is predetermined, concurrent speech can be constructed like a scripted monologue, though extraneous elements of speech (such as rate, dynamics, and stress) must be matched with others by monitoring their output and adjusting accordingly.

Concurrent Music: Concurrent music can be scripted or improvised, and furthermore, multiple people can simultaneously produce the same, or different, outputs. When concurrent performance is scripted, each individual can construct their contribution in a similar way to when performing individually, but extraneous elements of production (such as timing,

dynamics, and phrasing) are influenced by other performers. During improvisation, content can also be influenced by others.

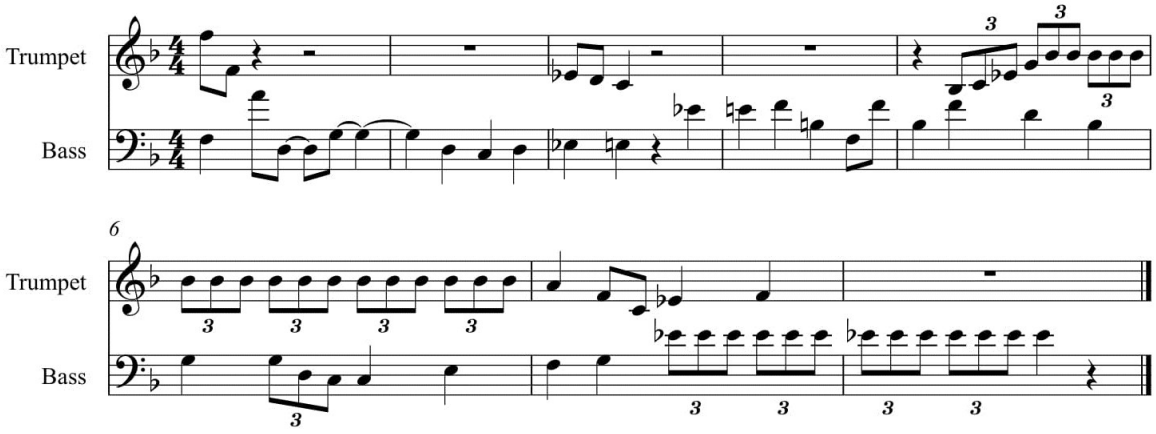

Figure 2. Example of concurrent music performance by two musicians, from Monson (1996).

In Figure 2, the top musical line is the performance of a trumpeter (Freddie Hubbard), and the bottom musical line is the performance of a bass player (Richard Davis). While the bassist in this sort of improvisation often holds a crucial role in establishing the harmony and rhythm by playing low single-beat notes, a clear break from this tendency is evident in bars 7-8. Here the bassist plays triplets (3 notes per beat) as opposed to a single note per beat, mimicking the pattern that was previously demonstrated by the trumpeter (bars 5-6). This passing of the triplet pattern between the two performers demonstrates the way two individuals can develop a musical theme together, and illustrates the challenge for interactors to consistently comprehend the output (timing and content) of others while planning and producing their own complementary output.

These examples have demonstrated that linguistic and musical interactions involve flexible adjustment based on the output of other interactors that do not arise when looking at

production or comprehension separately. We now focus on the cognitive and neural mechanisms underlying such shared cooperative activities in communicative interaction. We begin by proposing a framework for interaction, then outline its support in previous literature, and finally discuss a series of questions to be addressed in future work.

A Simulationist Account of Communicative Interaction

This account extends isolationist approaches to cover the potential simultaneity of production and comprehension (whether turns are organised concurrently or consecutively), and includes a means of making predictions across self and other. We draw on the model of Pickering and Garrod (2013), first focusing on individual production and comprehension, and then moving to what changes when these processes occur in different sorts of communicative interactions. According to Pickering and Garrod's (2013) theory of language, when a speaker generates a production command (i.e., an intention to communicate) they use a forward model to represent the predicted outcome of their speech before they finish producing it. The production command (comprising the linguistic message) is sent to the production implementer to initiate the utterance. At the same time, an efference copy of the production command is sent through a forward production model to generate a prediction of the upcoming output. This runs ahead of the actual production, and typically involves simplification of the production command. Speakers then compare what they produced with their predicted outcome (i.e., production percept), and any difference is fed back to update future predictions. This model improves the more it is used, as predictions are refined. Complementarily, listeners use an inverse model to predict what a speaker they are listening to will say next. Importantly, it is known that comprehenders interpret speech in a speaker-specific manner (Metzing & Brennan, 2003), and hence this inverse model takes any known differences between themselves and the speaker into account. Specifically, when the listener

covertly imitates the speech that they hear to derive the production command that would have been required to generate it, this transformation is based on the listener's production experience modified by known differences between themselves and the speaker (e.g. differences in vocal range). The listener then derives the production command for the upcoming speech using a forward model based on their own experience of talking modified by known differences between themselves and the speaker (e.g., different interests or speech habits). Similarly to individual production, the derived production command would lead to the generation of an expected percept, which is compared to the speech that they subsequently hear. Any difference would be fed back to improve the model and better tailor it to the specific speaker involved.

Garrod and Pickering (2015) argue that predictions of content and predictions of timing are separate. The content of the message can be represented at any level (phonological, syntactic, semantic), with different elements being emphasised in different situations. For example, if a speaker starts with 'I'm hungry, why don't we go to…' then their interlocutor would predict the semantic category of restaurant. However, only if the interlocutor knows the speaker's preferences could they also accurately predict the specific phonology of 'McDonalds'. The timing of the utterance, on the other hand, is incorporated into the prediction as context. Entrainment to the speaker's rate of speech (e.g. one syllable every 200ms) is used in combination with the content prediction to make a comprehensive prediction of the upcoming utterance (i.e. 3 syllables of 'McDonalds' x 200ms/syllable = 600ms to complete the utterance).

The simulation account of individual production and comprehension is also applicable to music, though music of course involves different representations. We follow many previous models of musical prediction to include representations of pitch and rhythm (Pearce & Wiggins, 2012). We define rhythm as the number/proportion of beats making up a note, and

again hold timing separately as beat-rate (i.e. tempo). To put this into a linguistic context, whereas words are a multiplication of syllables (and are dependent on the syllable-rate), musical rhythms are a multiplication of beats (and are dependent on the beat-rate). Note that differences in beat proportions or groupings are therefore a feature of rhythm, while variations in the placement of beats are a feature of timing (i.e. being early/late). Hence where the language model represents phonology, the music model represents pitch and rhythm. Furthermore, where the language model represents syntax, the music model represents the higher order organisation of pitch and rhythm, i.e., tonality and metre.

So far, the account of production and comprehension refers to monologue speaking or listening. However, interactions differ from individual production or comprehension in two ways: the dependence of consecutive outputs, and the interrelation of simultaneously occurring outputs. The dependence of consecutive outputs is based on the premise that cooperative shared activity involves mutual responsiveness, and hence that each individual can assume interdependence of contributions (i.e. that a partner's contribution will build on one's own contribution). The interrelation of simultaneously occurring outputs is based on the requirement to produce and comprehend at the same time; for example, when overlap occurs at turn-ends, when the listening individual interjects, or during concurrent turn production.

We propose that within an interaction, participants compute joint predictions, which involve a single integrated representation of both self and other. While the means of generating predictions are the same as in individual production or individual comprehension, an additional step is required to integrate these predictions into a single representation before they are compared with the resulting interaction output. Note that this process is proposed to underlie well-coordinated and successful interactions, when individuals are able to generate

accurate predictions about both themselves and their interaction partner; it may not occur when predictions are repeatedly violated.

Integration of self and other predictions occur both across consecutive contributions and during simultaneous contributions. In terms of integrating predictions of consecutive contributions, the cooperative nature of the interaction leads each individual to simulate their own contributions in relation to those of their partner as a high-level integrated representation. It is this integrated prediction that leads an interlocutor to expect that their question will be answered or that their antecedent phrase will be followed by a consequent phrase. It is also this integrated prediction that is violated if such events do not occur.

In terms of integrating predictions of simultaneously occurring contributions, the generation of a single joint prediction to use as a template against which to compare the joint output is a cognitively efficient means of monitoring interaction success. In concurrent turn organisation, i.e. choric language or a duet, the joint representation might include performance asynchrony as a difference between self and other (e.g. self being 50ms before other) as opposed to each individual's timing as an absolute value (e.g., self at 250ms and other at 300ms). In a duet the joint representation of content might also include pitch as a relationship between self and other, for example predicting a C major chord as opposed to oneself playing a C while a partner plays an E and a G. This joint prediction mechanism uses individual production and comprehension processes to generate individual predictions (comprehension processes again being tailored to the specific interactor), but then combines these predictions into a single joint prediction percept. Integrating the self and other predictions allows monitoring and repair more efficiently than multiple separate simultaneous predictions (i.e. one for each interactor) being first compared with each individual's output, and subsequently compared with each other to identify inter-individual mismatches. The reason for making a joint prediction (e.g., of a C major chord rather than a C, plus an E and a

G) is largely to reduce computational complexity (i.e., one prediction being made rather than two).

Importantly, however, there are times when people rely on individual rather than joint predictions when predicting simultaneously occurring contributions. Reliance on individual predictions may occur when joint predictions are often inaccurate, for example when the joint output includes frequent errors or inter-personal temporal instability. In such cases, identifying which part of the joint prediction was violated may be more complex than comparing each individual's prediction against each individual's output. Hence in such cases it would be more cognitively efficient to predict the individuals rather than the dyad. Even during one interaction, predictions may fluctuate between being joint and individual, potentially shifting from individual to joint as familiarity increases and predictions become more accurate, or shifting from joint to individual during periods of poor coordination. (Note that predictions across consecutive turns are always integrated, because each contribution is interdependent by virtue of being part of the shared cooperative interaction activity.)

Joint predictions relate to a behaviour (specifically, a joint behaviour) and are therefore "external" (relating to output). In addition, individuals make "internal" predictions about the sensorimotor experience of their own production. Notably, if sensorimotor predictions of one's own contribution are violated it may be possible to correct potential errors before they are output, which is not possible for "external" joint predictions (since the joint percept is required for monitoring). These internal predictions only relate to one's own output (because people do not have access to their partner's experience), and can provide additional information for identifying the agent of an error perceived in the joint output. For example, take the case of musicians playing a duet together at a fixed tempo, during which one musician comes in late. The joint prediction would be of simultaneous performance, and would be violated by the asynchrony of the entries. If this asynchrony were due to the first

musician's mistake, her internal prediction would also be violated, allowing attribution of the error to herself. On the other hand, if this asynchrony were due to her partner's mistake, her internal prediction would not have been violated. It is important to note that if the interaction constraints are unspecified, for example when playing a piece without a tempo marking, violations of joint predictions could alternatively indicate that interactors' assumptions of the overarching context mismatch. In such a situation both individuals may have fulfilled their internal predictions, but their different understandings of the context meant that their contributions produced an error when occurring in combination.

Context is the general knowledge that one interactor uses to inform their predictions across the interaction (i.e. including both interactors). In music, this could be an understanding of tempo or musical genre; in language, this could be an understanding of speech-rate or conversation topic. Notably, each interacting individual has their own representation of this overarching context, and of course these representations may be different. If there is evidence of mismatch between the outputs of two musicians that can be attributed to different understandings of general knowledge underlying the interaction, representation of the context can be adjusted appropriately. Representing the context thus allows efficient updating of parameters affecting prediction of the interaction as a whole. (We ignore the question of whether both interlocutors need to know that the knowledge is shared; see Clark, 1996).

Below we sketch a model focusing on how self and other representations are combined in interactive music and language. See Figure 3, illustrating the simplest and optimal case of the joint prediction mechanism for two interactors.

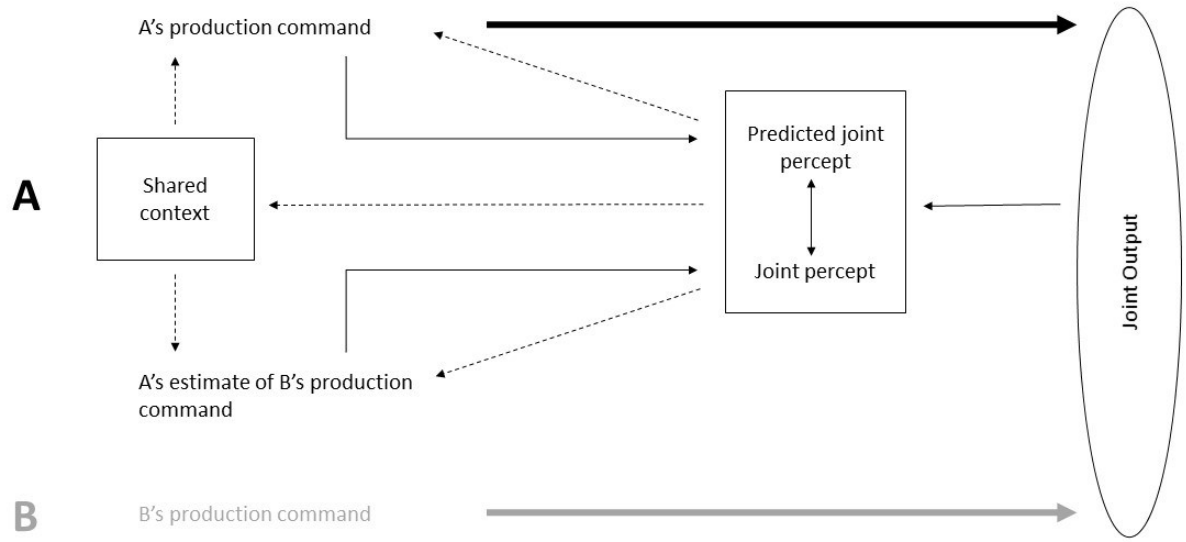

Figure 3. The processes used by Person A during an ideal interaction with Person B (involving music or language). In both music and language the production command is the intention underlying an individual's output (i.e. their produced speech or music performance). Person A produces her contribution (thick black uppermost arrow) while additionally predicting the contribution of Person B, with the full lines from production commands to the predicted joint percept indicating forward modelling of the production commands. The predicted joint percept is compared with the joint heard percept, and dotted lines indicate how feedback from this comparison is used to update the forward models (of either A or B) directly, or via context, in order to improve prediction accuracy and A's own outputs. Person B's contribution is shown in the thick grey lowermost arrow.

In this model, person A is both producing and comprehending (person B is doing the same but their full model is not shown). As described above, when A sends a production command she uses a forward model to predict the outcome of that production. Additionally, when she comprehends B she uses covert imitation and forward modelling to derive the production

command that B will use for his upcoming contribution. The means of generating these predictions are the same as for individual production and comprehension, but what differs in interaction is how the result of these (production and comprehension) predictions are compared with what is heard. Following the generation of these individual predictions, they are integrated into a single joint representation. Only when there is a mismatch between this prediction and the heard output (and one's internal prediction is not violated) would the joint prediction be decomposed into its constituent parts (i.e. self, other), to identify the locus of the error. Let's consider how this works in consecutive and concurrent turn organisation, each time from person A's perspective.

Consecutive interaction

Person A is engaged in a turn-taking interaction with Person B. Given the shared cooperative nature of the interaction, she expects that he will be responsive to her turns and that he is committed to interaction success. Hence while listening to person B, A covertly imitates him and predicts what he will produce next. As B is coming to the end of his turn and hence A has finished generating predictions about his output, she begins to generate the production commands for the start of her own contribution. Then as A's own contribution comes to an end, she predicts how B might respond. Though A makes predictions about her own and her partner's parts slightly differently (forward modelling for self-production, covert imitation followed by forward modelling for other-prediction), self and other predictions are integrated into a prediction of the interaction as a whole, based on the principles of shared cooperative activity. If responsiveness and coherence are not upheld, A may reconsider her categorisation of the situation as an interaction, and adjust her expectations to those of an individual monologue (e.g., in which interjections are not necessarily answered).

During each individual's turn, it is likely that the other provides a backchannel or supporting function. Note that this does not constitute a concurrent turn because of the asymmetry of content and relative salience of the contributions. Nonetheless, such simultaneous productions are also included in the integrated prediction of the consecutive turn situation, though these interjections may be simulated at a more abstract level than the turn itself (i.e. in language this simulation could be of an affirmation without specifying whether it will be 'yes' or 'right'; in music this simulation could be of a particular harmony without specifying the individual note sequence). If predictions of the supporting partner's output are relatively abstract, violations will occur only when the mismatch is critical to interaction success (e.g. if the interlocutor queries the speaker rather than agreeing, or the accompanist produces a harmonic/metric clash).

Notably, if A and B are interacting smoothly, there should be little difference between A's joint prediction and the actual joint output. On the other hand, if A and B are not interacting smoothly, A's joint prediction will often clash with the actual joint output, leading to different parts of the model being updated depending on the reason for the error.

Consecutive Language: To demonstrate how an individual would predict a conversation partner, we refer to a real example. In the discussion between a public relations officer and an engineer following a hurricane in Figure 4, Person A gives a request for information at (1) which B appears to respond to adequately at (2). However, during (2) an error is signalled by A's monitor because while she predicted that both herself and her partner would use the term *ones* to refer to roads, the joint output shows that B actually used *ones* to refer to shelters. The misunderstanding of the referent of *ones* is therefore fed back to update A's representation of the context (i.e. the general knowledge that she uses to inform her

predictions of herself and B), which leads her to update her upcoming production command to allow clarification at (3).

1 - **A:** Which ones are closed, an' which ones are open

2 - **B:** (pointing to map) Most of 'em. This, this, this, this

3 - **A:** I don't mean on the shelters, I mean on the roads

4 - **B:** Oh!

Figure 4. Simplified conversation example from Schegloff (1987).

Consecutive Music: In music involving turn-taking, many of the same issues from conversation occur. In an improvisation between A and B, A would play her turn and when coming towards the end, begin to predict B's response. If B doesn't respond as anticipated, for example by introducing a new theme rather than providing a variation of A's theme, A may update her model of his improvisation style (i.e. A's understanding of the differences between herself and B). Alternatively, if B does provide a variation of her theme but unexpectedly modulates to a different key, A may not only update her understanding of differences between herself and B (in terms of harmonic style), and also her representation of the context (i.e., updating the musical key as this will affect both his and her own upcoming contributions).

Importantly, B may interject during A's turn. A could include this accompaniment (somewhat akin to linguistic backchanneling) in her joint prediction. She may predict that B will interject with the harmonic progression that she is performing to emphasise specific beats. Similarly, she can make predictions about her own output when she accompanies B's turn. Hence there are points that she will anticipate a simultaneous output, even though the turn structure is consecutive.

Notably, the explicit beat structure of music does lead to some differences between music and language interaction. Unlike language, it is rare to correct a musical error (i.e. attempting the note again), because this would disrupt the musical flow and may shift the temporal pattern of the melody to the wrong beat. So while in certain cases self-corrections can be made during performance (e.g. if a note on the violin is flat, it can be sharpened while it is being produced), in many cases errors are left uncorrected with commands for upcoming content being adjusted instead.

Concurrent interaction

In this situation, both individuals concurrently produce and listen to their combined output at all times. Person A predicts how her upcoming contribution will feel to produce (internal prediction), as well as how her own and B's contribution will sound in a single representation of the joint output (external prediction). As in consecutive production, A's prediction of B takes into account known differences between the two individuals. Any violations of this joint representation are fed back to update the relevant part of the model.

Concurrent Language: When chanting at a football match, individuals produce the same content at the same time 'We're going to win, 3-nil, 3-nil'. However, there may be interpersonal variation in content, for example some fans being more optimistic than others ('We're going to win, 5-nil, 5-nil'). Assuming that each individual predicts that their version of the chant will be spoken by all, such differences would lead to an error signal between the predicted joint percept and the actual joint percept. While too late to alter the words already spoken, this error signal could be used to update the representation of the context (e.g., to converge on '4-nil, 4-nil') and hence alter one's production commands for the upcoming repeat. Aside from content (which in most cases should be the same during concurrent turn production in language), individuals monitor for mismatch in terms of temporal, dynamic, or

other acoustic features of the sound. For example, in terms of timing, if B begins the word 'win' earlier than A, A would recognise an error between her prediction of a simultaneous joint utterance, and the actual output, and could update her representation of speech-rate in her context representation. It is important to note, however, that the way that such a mismatch between a joint prediction and joint percept is dealt with may vary depending on the role of the individual. When each interactor holds an equal role in the interaction they may both update their representations of context to converge on a compromise between their starting tempi, but when each interactors holds a different role - A being the coach and B being a fan - A may continue without adjustment while B adjusts to match A.

Concurrent Music: If performers A and B are playing monophonically (i.e. they play the same notes at the same time), then each performer's joint prediction would be for a pair of notes to occur simultaneously, at the same pitch and for the same duration. A would predict her own part using her forward model of herself, and predict B's part using her forward model of B taking into account knowledge of differences between herself and B. For example, if A knew that B was less skilled than her, she may use a subset of her performance experience to predict B's upcoming contribution. In this case she may rely on her own performance experience from when she was less skilled, in order to identify potential times of difficulty for B, and prepare her own contribution sympathetically (i.e. wait longer to produce her note if she anticipates B will have difficulty finding his).

In a smooth interaction, the notes of A and B would occur at the same time (i.e. within bounds of acceptable temporal synchrony), and no error would be recorded by the monitor. But if there is a temporal error, an error would be elicited. For example, A and B may be playing a piece with a predetermined tempo of one beat every 500ms (i.e., this is the tempo marking on the score). If A played note 2 at timepoint 500ms but B played it at 600ms, several of A's joint predictions would be violated: her prediction of two simultaneous notes

(joint prediction), and her prediction that the two notes would both be based on a beat lasting 500ms. At this point, A would disambiguate the difference between her predicted joint percept and the actual joint percept according to whose contribution was whose (referring to her internal model as well), and the monitor would identify the error being between A's prediction for B (that B would play at the 500ms timepoint), and B's output. Given that the tempo had been previously specified, A would identify B's performance as an error (this tempo is represented in A's context representation), and would continue playing 'correctly' (without adjusting her performance), assuming B would adjust his performance appropriately. A similar thing could happen if the error were pitch-based in a pre-composed piece (i.e. if B played out of key): again, A could continue and expect B to fix his own errors. If, however, the tempo or notes of the piece could be negotiated between musicians (i.e. in a piece with no specific tempo, or an improvisation), any differences between musicians could be handled differently. In the piece without a predetermined tempo, A may instead take the difference between her predicted tempo and person B's starting tempo to update her model of context, for example altering her representation of beat duration from 500ms per beat to 550ms per beat. She would use this to adjust her subsequent performance commands and predictions of B. However, which musician(s) adjust may again be based on roles within the interaction (primary/accompanying instrument), and would result in the generation of leader/follower relationships. The closer A and B's representations of context become, the smoother the interaction, as interpersonal predictions become more accurate.

Other features impacting interaction

We propose that this model spans both scripted and improvised interactions, and familiar and unfamiliar partners, but do not claim that contributions are predicted in the same way between contexts. When interactions are scripted, and the script for each interactor is known

by both, the joint prediction will be highly specified for both individuals. In such a case, A may predict the stylistic aspects (such as accent, dynamics, or temporal variability) of B's output as well as the specific words/notes he will produce, though the accuracy of these predictions will depend on A's familiarity with B's style. When B is familiar and A can use a well-tuned model to predict B's output, then A's predictions will be highly accurate. When B is unfamiliar, on the other hand, and A cannot use a well-tuned model to predict B's output, A's predictions are instead based on her own idiosyncratic model, which may turn out to be highly inaccurate.

Hence knowing the content of a partner's contribution does not necessarily improve fine-grained temporal prediction, although it may improve larger-scale predictions based on the linguistic or musical structure (Ragert, Schroeder, & Keller, 2013). Instead fine-grained temporal predictions improve as the partner's style becomes more familiar and one's model of that partner becomes better calibrated (Ragert et al., 2013). In a similar way, improvised interactions may involve less specific, or more flexible, predictions (Bianco, Novembre, Keller, Villringer, & Sammler, 2018), and these may improve through familiarity with a partner's improvisation style.

Support for the Model

The three main claims of our model presented above are:

1) Individuals generate a single prediction percept combining both interactors
2) Predictions of other interactors are made via simulation
3) Similar representations of context between interactors improves coordination

Evidence for these claims will be discussed in turn.

1. Do we generate a single prediction percept combining both interactors?

The model proposes that in cooperative joint actions we don't just generate two individual prediction percepts – one 'self' and one 'other' – but instead generate a single combined prediction percept that is based on the contributions of both interactors. While there is little research that directly addresses this issue, there are reasons to support this proposal. Initial evidence comes from individuals representing the contribution of their partners as well as themselves, a premise that has support across joint action research (including work on language). Further evidence for this claim comes from individuals prioritising the shared goal of an interaction over either individual's goal.

A variety of joint action paradigms have been used to show co-representation of a partner. Many studies have demonstrated interference between one's own and another individual's action using the joint Simon task, in which the standard Simon task is shared across two people (each taking one response button) (Dolk et al., 2014). In the Simon task, two types of target stimuli are presented on the left or the right side of a computer screen. One type of target should be responded to with the left hand, and the other type of target should be responded to with the right. When a target appears on the same side of the screen as its response mapping, performance is facilitated. When the target appears on the opposite side as its response mapping, performance is impaired. In the joint Simon task, two participants sit next to each other, each responding to different types of target (each with only one hand). Sitting next to an individual responding to the 'other' cue elicits behavioural and neural interference in a similar way to when responding to both cues oneself (Sebanz, Knoblich, & Prinz, 2003), indicating that both individuals' actions are represented.

There is also evidence that prediction about one's own and another's actions are combined in naturalistic situations. A study investigating lifting and balancing found that a participant find it easier to keep a tray balanced while lifting a glass from someone else's tray when that other person simultaneously, rather than sequentially, lifts a glass from the participant's own tray

(Pezzulo, Iodice, Donnarumma, Dindo, & Knoblich, 2017). This suggests that representing one's own action is easily translated to another, and facilitates one's response to that other individual's action. Furthermore, in a study of lifting and clinking glasses (Kourtis, Knoblich, Woźniak, & Sebanz, 2014), when lifting a glass to the centre of the table, participants showed greater neural activation when they were to clink with another person than when they were acting alone, demonstrating the influence of predicting others on one's own action representation.

There is also evidence of co-representation of a partner's speech. Individuals take longer to name pictures when they believe that both themselves and their partner are naming pictures at the same time (Gambi, Van de Cavey, & Pickering, 2015). The fact that the partner's task interfered with the participant's own suggests that individuals represent the speech of their partners during an interactive task. Similar effects of a partner's speech on one's own have been demonstrated in a semantic interference task (Kuhlen & Rahman, 2017), and neural evidence for co-representing partner in a go/no-go picture naming task indicates lexical processing for both one's own and a partner's spoken response (Baus et al., 2014).

While in both music and language it is difficult to differentiate a joint prediction from multiple individual predictions, concurrent production allows the issue to be addressed directly. Turning to research in musical interaction, it has been found that when musicians know both parts of a (concurrent) duet piece, they adjust the temporal profile of their own performance according to the complexity of their partner's part (Loehr & Palmer, 2011), suggesting a focus on the joint outcome. Furthermore, when feedback is manipulated in duetting musicians' parts that either violate only their individual goals (i.e. to perform their own part correctly) or additionally violate their shared goal (i.e. to perform a harmonically coherent musical piece), neural responses differ (Loehr, Kourtis, Vesper, Sebanz, & Knoblich, 2013). The P300, indicating updating of context (Donchin & Coles, 1988) is

enhanced when the shared goal (i.e. the harmonic structure of the piece) is violated compared to when either individual's goal is violated. This finding indicates prediction of the joint outcome, i.e. that both musical lines are integrated into a joint prediction of the harmonic pattern that they make in combination, with the mismatch between this joint prediction and the real percept eliciting disruption.

Indeed, concurrent music performance (in which musicians play different musical lines) offers the possibility of attending to a combined interaction outcome, but it is also possible to prioritise individual contributions within this combined output. Keller's Prioritised Integrative Attending theory specifies that attentional weighting can vary in such a situations, with prioritisation of one's own part within such a performance being optimal (Keller, 1999). An elegant study tested how interpersonal entrainment accuracy in a piano duet affected self-other integration (Novembre, Sammler, & Keller, 2016). Pairs of pianists played duets while alpha activity, hypothesised to be involved in integrating representations of self and other, was recorded. Importantly, alpha suppression has been linked to increased cognitive processing efficiency (Klimesch, 2012), and hence was used as a measure of successful self-other integration. When musicians were well entrained and performing together synchronously, alpha activity was indeed suppressed, indicating integration of self-other predictions. However, when musicians were poorly entrained and performing together asynchronously, alpha activity was enhanced, indicating reduced integration of self-other predictions. This dissociation supports our proposal that when joint predictions are frequently erroneous (such as during the asynchronous performance), each individual's predictions are held separate.

2. Do people make predictions about interactors via simulation?

There are strong indications that people indeed use simulation to coordinate with a partner during interaction (though often this has been tested with non-adaptive interaction partners).

Evidence for the use of simulation in language comes from a joint picture naming task (Gambi, Cop, & Pickering, 2015). Participants took turns to name images that were occasionally replaced by a different target picture during naming. The target picture signalled that they should stop naming the original image, and either that they, their partner, or neither, should name the new image. Strikingly, the original speaker was more likely to finish naming the original picture regardless of whether it was their role to name the target or their partner's, compared to when it was neither of their roles. Such interference suggests motor activation occurs similarly for both one's own speech and that of an interaction partner. Furthermore, in a go/no-go picture naming task that varied whether the participant or a confederate should respond, the participants showed greater inhibition during the confederate's responses, indicating motor system involvement (Baus et al., 2014).

Moreover, evidence from music suggests that our predictions of others involve simulation (Novembre, Ticini, Schutz-Bosbach, & Keller, 2012), and that these simulations are based on the idiosyncrasies of our own motor performance. For example, pianists are better at detecting perturbations in their own performances than those of others (Repp & Keller, 2010), and are also better at synchronising with a previous recording of themselves than a previous recording of another pianist (Keller et al., 2007). Furthermore, while we might expect pianists to benefit from familiarity with the partner's part in a duet, when pianists have practiced the other part of a piano duet themselves they are actually worse at subsequently synchronising with another pianist playing it (Ragert et al., 2013). This suggests that they use their own experience to predict their duet partner, and without knowledge of how they and their partner differ, these erroneous predictions lead to impaired synchronisation.

The more similar we are to somebody, the better we are at synchronising with them. Individuals with similar spontaneous performance rates are able to synchronise more accurately than individuals with dissimilar spontaneous performance rates (Loehr & Palmer,

2011; Zamm, Pfordresher, & Palmer, 2015), with performance asynchrony being correlated with the difference in the spontaneous rates chosen by the performers. In other words, even during attempts at coordination, spontaneous rate affects asynchrony. Furthermore, musicians with similar performance rates show mutual adaptation in duet performances, as opposed to falling into a leader-follower-type relationship (in which one performer dominates, Loehr & Palmer, 2011). These findings are consistent with predictions being based on one's own motor experience, because when interactors are more similar their predictions will be more accurate and adjustments more beneficial.

Moreover, two recent studies of musical duets showed a causal role of simulation for coordination accuracy, both for consecutive (Hadley et al., 2015) and concurrent (Giacomo Novembre, Ticini, Schütz-Bosbach, & Keller, 2014) turn organisation. In these studies, musicians were given either both parts of a duet, or only their own part of a duet, to memorise. One part of the duet was played with the right hand, one with the left. They then came into the lab and played only one part of each duet in time with an audiovisual recording of another pianist. Applying double pulse TMS to motor regions relating to the participant's unused hand (i.e. the hand that the partner was using) either around a tempo change, or a turn-switch, caused performance disruption for the participant's own part only when they had previously practiced their partner's part. The motor regions targeted had been implicated in previous research on simulation (Lahav, Saltzman, & Schlaug, 2007), and so it appears that such simulation is involved in coordinating with a partner. More specifically, when playing with a partner whose actions are strongly encoded in one's own motor system (due to direct experience of producing their part), this simulation mechanism (involving the dorsal premotor cortex) is causally involved in accurate coordination.

3. Do similar context representations between interactors improve coordination?

In speech, context includes one's general knowledge about the content; in music, context includes one's understanding of musical features such as harmony and rhythm. In both speech and music, context includes one's understanding of rate or tempo. We previously suggested that similar understandings of context would facilitate interaction due to interpersonal predictions becoming more accurate. Indeed when two musicians have similar representations of context in terms of upcoming tempo changes, they play more synchronously even in the preceding phrases (Novembre et al., 2016). Furthermore, simulation is enhanced during observation of a partner when more information about context is available; pianists show enhanced responses when observing mute piano performances including incongruent musical progressions in 5-chord sequences compared to 2-chord sequences (Sammler et al., 2013). As it is also evident that interacting individuals align on features such as rate of production over time (Cohen Priva, Edelist, & Gleason, 2017; Finlayson, Lickley, & Corley, 2012; Jungers & Hupp, 2009; Schultz et al., 2016), we speculate that successful interaction results from a predictive simulation mechanism that makes use of interactors' representations of context (Colling & Williamson, 2014). Furthermore, we propose neural synchrony as a neural marker of similar context representations.

Neural coupling between interacting individuals has been demonstrated in both linguistic and musical interactions. In a study of conversation that recorded neural activity from both interactors, oscillations synchronised between the speaker and the listener during turn-taking speech (Pérez, Carreiras, & Duñabeitia, 2017), while in concurrent music performance, musicians show increased phase synchronisation around the onset of a music performance when cued by a metronome (Lindenberger, Li, Gruber, & Müller, 2009), even when playing

different musical lines (i.e. polyphonically; Sänger, Müller, & Lindenberger, 2013). Notably, such interpersonal entrainment is influenced by the relationship between the interactors. In a musical interaction, neural coupling at 12Hz from the leader to the follower but not from follower to leader, is greater after, compared to before, note onsets (Sänger et al., 2013), potentially indicating that followers adjust their representation of context to the leader following each musical event.

If similar understanding of context contributes to interaction success, there should be a relationship between neural synchronisation and behavioural measures reflecting this success. Indeed, neural coupling between speakers and listeners relates to their ability to make predictions about what will come next (Dikker, Silbert, Hasson, & Zevin, 2014) and to the level of comprehension during the interaction (Stephens, Silbert, & Hasson, 2010). Furthermore, a study of speech rhythm showed that the amplitude of entrained alpha and theta activity correlates with interpersonal synchronisation accuracy (Kawasaki, Yamada, Ushiku, Miyauchi, & Yamaguchi, 2013), and in music, phase alignment between the brains of interacting musicians correlates with their play-onset synchrony (Lindenberger et al., 2009).

Strong evidence of a causal relationship between synchronisation of neural processes and interaction success comes from a recent transcranial stimulation study investigating finger tapping. Novembre and colleagues (Novembre, Knoblich, Dunne, & Keller, 2017) induced synchronised or non-synchronised beta oscillations between pairs of individuals engaged in a rhythmic finger-tapping task using transcranial alternating current stimulation. When stimulation induced synchronised beta activity over the left centro-parietal area, participants' first few finger taps were more synchronised than when they experienced sham stimulation. Accuracy did not improve when unsynchronised beta activity was induced, or when synchronised alpha activity was induced, indicating a specific role of beta synchronisation in

interpersonal coordination. Further study of the relationship between beta synchronisation and interaction success, particularly in relation to tempo in concurrent music or language production, could elucidate the role of this phenomenon.

Conclusions and Future Directions

We have presented a framework for communicative interaction that specifies a means by which interactors can share representations, predict actions, and integrate predicted effects of their own and others' actions - the fundamental requirements for performing joint action (Keller, 2014; Sebanz, Bekkering, & Knoblich, 2006). This simulationist model provides the novel possibility of predicting multiple interactors' contributions jointly (as a single percept), and furthermore implements a context component to efficiently update parameters of the interaction that affect all involved in a single step. This model covers how individuals can both establish and maintain synchrony, and is speculatively linked to recent literature on neural synchronisation during interaction.

It is now necessary to test the hypotheses relating to this model, and situations in which it applies. Most notably, this model proposes each interactor predicts one joint percept including the contributions of both interactors. Perhaps the simplest test of such a proposal is how individuals deal with a violation of the predicted percept when it is due to one individual's error, or both individuals' errors. If a joint percept is predicted (as we propose), it should not matter if the error is due to one or both individuals, but if two individual percepts are predicted, errors committed by two individuals would cause more disruption than errors committed by one. Importantly, internal predictions would need to be considered, and matched across conditions. This would mean comparing the response to one's own error with the response to one's own plus a partner's error.

In addition, the model proposes that similar context representations are causally involved in interaction success. Evidence to support this supposition was drawn from studies indicating that synchronised neural oscillations relate to temporal coordination in finger tapping, music, and speech studies. However, neural coupling (and behavioural entrainment) can also be understood in terms of dynamic systems theory (Schmidt, Carello, & Turvey, 1990), which proposes that interacting systems become aligned over time due to general laws of attraction. A potential way to test whether similar neural signatures indicate similar understandings of context would be to measure neural activity prior to interaction when interactors are given the same or different contextual information (presumably relating to tempo). If neural oscillations are used proactively, then neural activity between individuals with the same understanding of context should be more strongly synchronised before interaction begins than individuals with different understandings of context. In addition, novel paradigms should be developed to identify whether neural synchronisation occurs predictively for contextual information such as semantic understanding.

Finally, the sorts of interaction that this model serves should be tested. We have focused on shared cooperative activity (Bratman, 1992), as this is the simplest form of linguistic and musical interaction, but there has recently been a wealth of recent research into intra-personal synchronisation in cooperation vs competition (Balconi & Vanutelli, 2017). Since individuals in a cooperative context show greater synchronisation than those in a competitive context (Cui, Bryant, & Reiss, 2012), whether the model we have presented is specific to cooperative activity should be tested. Finally, we did not consider how the model scales up to larger numbers of interactors. It is common to chat or chant in a group, or play music in an orchestra, but when predicting the output of multiple others does one simulate the average of the group, the most dominant individual, or each individual separately?

We have argued that comparing musical and linguistic interactions allows basic prediction mechanisms to be investigated in an ecologically valid yet highly controlled manner. We have presented a framework providing numerous testable hypotheses for future investigation, in order to advance understanding of communicative joint action and improve understanding of interpersonal prediction more generally. We intend the framework that we have presented to assist with this exploration of the mechanisms of interaction.

References

Balconi, M., & Vanutelli, M. E. (2017). Cooperation and Competition with Hyperscanning Methods: Review and Future Application to Emotion Domain. *Frontiers in Computational Neuroscience*, *11*, 86.

Baus, C., Sebanz, N., Fuente, V. de la, Branzi, F., Cognition, C. M.-, & 2014, U. (2014). On predicting others' words: Electrophysiological evidence of prediction in speech production. *Cognition*, *133*(2), 395–407.

Becchio, C., Sartori, L., & Castiello, U. (2010). Toward You: The Social Side of Actions. *Current Directions in Psychological Science*, *19*(3), 183–188.

Bianco, R., Novembre, G., Keller, P. E., Villringer, A., & Sammler, D. (2018). Musical genre-dependent behavioural and EEG signatures of action planning. A comparison between classical and jazz pianists. *NeuroImage*, *169*, 383–394.

Bögels, S., Barr, D. J., Garrod, S., & Kessler, K. (2015). Conversational Interaction in the Scanner: Mentalizing during Language Processing as Revealed by MEG. *Cerebral Cortex*, *25*(9), 3219–3234.

Bratman, M. (1992). Shared cooperative activity. *The Philosophical Review*, *101*(2), 327–341.

Clark, H. (1996). *Using language*. Cambridge University Press.

Cohen Priva, U., Edelist, L., & Gleason, E. (2017). Converging to the baseline: Corpus

evidence for convergence in speech rate to interlocutor's baseline. *The Journal of the Acoustical Society of America*, *141*(5), 2989-2996.

Colling, L. J., & Williamson, K. (2014). Entrainment and motor emulation approaches to joint action: Alternatives or complementary approaches? *Frontiers in Human Neuroscience*, *8*, 754.

Cui, X., Bryant, D. M., & Reiss, A. L. (2012). NIRS-based hyperscanning reveals increased interpersonal coherence in superior frontal cortex during cooperation. *NeuroImage*, *59*(3), 2430–2437.

Cummins, F. (2003). Practice and performance in speech produced synchronously. *Journal of Phonetics*, *31*(2), 139–148.

Dikker, S., Silbert, L. J., Hasson, U., & Zevin, J. D. (2014). On the same wavelength: predictable language enhances speaker-listener brain-to-brain synchrony in posterior superior temporal gyrus. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *34*(18), 6267–6272.

Dolk, T., Hommel, B., Colzato, L. S., Schutz-Bosbach, S., Prinz, W., & Liepelt, R. (2014). The joint Simon effect: a review and theoretical integration. *Frontiers in Psychology*, *5*, 974.

Donchin, E., & Coles, M. G. H. (1988). Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences*, *11*(03), 357.

Finlayson, I., Lickley, R., & Corley, M. (2012). Convergence of speech rate: Interactive alignment beyond representation. In *Twenty-Fifth Annual CUNY Conference on Human Sentence Processing*.

Fitch, W. (2006). The biology and evolution of music: A comparative perspective. *Cognition*, *100*(1), 173–215.

Fuchs, T., & De Jaegher, H. (2009). Enactive intersubjectivity: Participatory sense-making

and mutual incorporation. *Phenomenology and the Cognitive Sciences*, *8*(4), 465–486.

Gambi, C., Cop, U., & Pickering, M. (2015). How do speakers coordinate? Evidence for prediction in a joint word-replacement task. *Cortex*, *68*, 111–128.

Gambi, C., Van de Cavey, J., & Pickering, M. (2013). A joint interference effect in picture naming. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*.

Garrod, S., & Pickering, M. (2015). The use of content and timing to predict turn transitions. *Frontiers in Psychology*, *6*.

Jungers, M. K., & Hupp, J. M. (2009). Speech priming: Evidence for rate persistence in unscripted speech. *Language and Cognitive Processes*, *24*(4), 611–624.

Kawasaki, M., Yamada, Y., Ushiku, Y., Miyauchi, E., & Yamaguchi, Y. (2013). Inter-brain synchronization during coordination of speech rhythm in human-to-human social interaction. *Scientific Reports*, *3*(1), 1692.

Keller, P. (1999). Attending in Complex Musical Interactions: The Adaptive Dual Role of Meter. *Australian Journal of Psychology*, *51*(3), 166–175.

Keller, P. E. (2014). Ensemble percormance: Interpersonal alignment of musical expression. In D. Fabian, R. Timmers, & E. Schubert (Eds.), *Expressiveness in music performance: Empirical approaches across styles and cultures* (pp. 260–282).

Keller, P. E., Knoblich, G., & Repp, B. H. (2007). Pianists duet better when they play with themselves: On the possible role of action simulation in synchronization. *Consciousness and Cognition*, *16*(1), 102–111.

Klimesch, W. (2012). Alpha-band oscillations, attention, and controlled access to stored information. *Trends in Cognitive Sciences*, *16*(12), 606–617.

Kourtis, D., Knoblich, G., Woźniak, M., & Sebanz, N. (2014). Attention Allocation and Task Representation during Joint Action Planning. *Journal of Cognitive Neuroscience*, *26*(10), 2275–2286.

Kuhlen, A. K., & Rahman, R. (2017). Having a task partner affects lexical retrieval: Spoken word production in shared task settings. *Cognition*, *166*, 94–106.

Lahav, A., Saltzman, E., & Schlaug, G. (2007). Action representation of sound: audiomotor recognition network while listening to newly acquired actions. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *27*(2), 308–314.

Lindenberger, U., Li, S.-C., Gruber, W., & Müller, V. (2009). Brains swinging in concert: cortical phase synchronization while playing guitar. *BMC Neuroscience*, *10*(1), 22.

Loehr, J. D., Kourtis, D., Vesper, C., Sebanz, N., & Knoblich, G. (2013). Monitoring Individual and Joint Action Outcomes in Duet Music Performance. *Journal of Cognitive Neuroscience*, *25*(7), 1049–1061.

Loehr, J. D., & Palmer, C. (2011). Temporal Coordination between Performing Musicians. *Quarterly Journal of Experimental Psychology*, *64*(11), 2153–2167.

Menenti, L., Pickering, M. J., & Garrod, S. C. (2012). Toward a neural basis of interactive alignment in conversation. *Frontiers in Human Neuroscience*, *6*, 185.

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, *49*(2), 201–213.

Monson, I. T. (Ingrid T. (1996). *Saying something : jazz improvisation and interaction*. University of Chicago Press.

Noordzij, M. L., Newman-Norlund, S. E., De Ruiter, J. P., Hagoort, P., Levinson, S. C., & Toni, I. (2009). Brain mechanisms underlying human communication. *Frontiers in Human Neuroscience*, *3*, 14.

Novembre, G., Knoblich, G., Dunne, L., & Keller, P. E. (2017). Interpersonal synchrony enhanced through 20 Hz phase-coupled dual brain stimulation. *Social Cognitive and Affective Neuroscience*, *12*(4), nsw172.

Novembre, G., Sammler, D., & Keller, P. E. (2016). Neural alpha oscillations index the balance between self-other integration and segregation in real-time joint action. *Neuropsychologia*, *89*, 414–425.

Novembre, G., Ticini, L. F., Schutz-Bosbach, S., & Keller, P. E. (2012). Distinguishing Self and Other in Joint Action. Evidence from a Musical Paradigm. *Cerebral Cortex*, *22*(12), 2894–2903.

Novembre, G., Ticini, L. F., Schütz-Bosbach, S., & Keller, P. E. (2014). Motor simulation and the coordination of self and other in real-time joint action. *Social Cognitive and Affective Neuroscience*, *9*(8), 1062–1068.

Patel, A. D. (2003). Language, music, syntax and the brain. *Nature Neuroscience*, *6*(7), 674–681.

Pearce, M. T., & Wiggins, G. A. (2012). Auditory Expectation: The Information Dynamics of Music Perception and Cognition. *Topics in Cognitive Science*, *4*(4), 625–652.

Pérez, A., Carreiras, M., & Duñabeitia, J. A. (2017). Brain-to-brain entrainment: EEG interbrain synchronization while speaking and listening. *Scientific Reports*, *7*(1), 4190.

Pezzulo, G., Iodice, P., Donnarumma, F., Dindo, H., & Knoblich, G. (2017). Avoiding Accidents at the Champagne Reception. *Psychological Science*, *28*(3), 338–345.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(04), 329–347.

Ragert, M., Schroeder, T., & Keller, P. E. (2013). Knowing too little or too much: the effects of familiarity with a co-performer's part on interpersonal coordination in musical ensembles. *Frontiers in Psychology*, *4*, 368.

Repp, B. H., & Keller, P. E. (2010). Self versus other in piano performance: detectability of timing perturbations depends on personal playing style. *Experimental Brain Research*, *202*(1), 101–110.

Sacks, H., Schegloff, E., & Jefferson, G. (1974). A Simplest Systematics for the Organization of Turn-Taking for. *Language*, *50*, 696–735.

Sammler, D., Novembre, G., Koelsch, S., & Keller, P. E. (2013). Syntax in a pianist's hand: ERP signatures of "embodied" syntax processing in music. *Cortex*, *49*(5), 1325-1339.Sänger, J., Müller, V., & Lindenberger, U. (2013). Directionality in hyperbrain networks discriminates between leaders and followers in guitar duets. *Frontiers in Human Neuroscience*, *7*, 234.

Schegloff, E. (1987). Some sources of misunderstanding in talk-in-interaction. *Linguistics*, *25*(1), 201–218.

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience. *Behavioral and Brain Sciences*, *36*(04), 393–414.

Schmidt, R., Carello, C., & Turvey, M. (1990). Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(2), 227–247.

Schultz, B., O'Brien, I., Phillips, N., McFarland, D., Titone, D., & Palmer, C. (2016). Speech rates converge in scripted turn-taking conversations. *Applied Psycholinguistics*, *37*(05), 1201–1220.

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends in Cognitive Sciences*, *10*(2), 70–76.

Sebanz, N., Knoblich, G., & Prinz, W. (2003). Representing others' actions: just like one's own? *Cognition*, *88*(3), B11–B21.

Stephens, G. J., Silbert, L. J., & Hasson, U. (2010). Speaker-listener neural coupling underlies successful communication. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(32), 14425–14430.

Stivers, T., Enfield, N., Brown, P., Englert, C., Hayashi, M., Heinemann, T., … Levinson, S. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(26), 10587–10592.

Zamm, A., Pfordresher, P. Q., & Palmer, C. (2015). Temporal coordination in joint music performance: effects of endogenous rhythms and auditory feedback. *Experimental Brain Research*, *233*(2), 607–615.